Мягкая оценка согласия ансамблевых языковых моделей при автоматической разметке текстовых данных

И. Р. Мусин

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

im_rasulev@vk.com

Аннотация. **В работе рассматривается задача автоматической разметки новостных текстов с использованием ансамбля из четырёх больших языковых моделей. Предложена метрика agreement rate — мягкая мера согласия ансамбля, определяемая через нормированное среднее абсолютное отклонение тональных оценок. Разработана и апробирована система ELNA (Ensemble LLM News Analyzer), обеспечивающая классификацию текстов по шести тематическим категориям, четырём типам срочности и числовую оценку тональности. Экспериментальная верификация проведена на корпусе из 5 702 криптовалютных и финансовых новостей. Средний уровень межмодельного согласия составил 0,942 при стандартном отклонении 0,035. Выявлена умеренная положительная корреляция метрики согласия с абсолютным значением тональности, интерпретируемая как следствие полярности эмоциональных оценок. Предложен практический критерий качества разметки: документы с низким уровнем согласия рекомендуются к ручной экспертной верификации.**

Ключевые слова: большие языковые модели, ансамблевое обучение, мягкие измерения, согласие моделей, автоматическая разметка текста, обработка естественного языка, анализ тональности

I. Введение

Автоматическая разметка текстовых корпусов является одной из базовых задач обработки естественного языка и приобретает особую значимость в условиях непрерывно растущих объёмов информации [7]. В финансовой сфере актуальность задачи определяется необходимостью оперативного анализа новостного фона для систем поддержки принятия торговых решений.

Широкое распространение больших языковых моделей (LLM) открыло возможность перехода от ручной разметки к малому числу обучающих примеров (few-shot) [4]. Однако отдельная LLM не является детерминированным классификатором: её ответ существенно зависит от архитектуры, предобучения и формулировки запроса [2]. Это порождает задачу оценки надёжности автоматической разметки.

Ансамблевые методы давно доказали свою эффективность в задачах классификации [6]: объединение независимых классификаторов снижает дисперсию ошибки. Перенос этого приема на ансамбли современных языковых моделей — вполне естественный шаг [9, 1].

Цель работы разработать и верифицировать мягкую метрику межмодельного согласия ансамбля LLM применительно к задаче тематической и тональной разметки новостных текстов.

Задачи:

1. Разработать формализованную метрику AR как меру неопределённости ансамблевой классификации.
2. Реализовать параллельный конвейер мультиагентной разметки на реальном корпусе.
3. Исследовать зависимость AR от тематики, типа срочности и оценки значимости.

II. Методология

Система ELNA (Ensemble LLM News Analyzer) реализует параллельный конвейер обработки, в котором каждый документ независимо анализируется четырьмя LLM-агентами. Финальные метки формируются посредством агрегации индивидуальных голосов, как показано на рис. 1.

Надёжность конвейера обеспечивается двумя механизмами. Каждый запрос к агенту ограничен таймаутом в 45 секунд: агент, не ответивший в срок, исключается из агрегации для данного документа, тогда как остальные продолжают формировать консенсус. При трёх последовательных ошибках агент деактивируется на всю сессию, что предотвращает накопление задержек при исчерпании API-лимитов, именно этим объясняется высокий процент ошибок у Grok 4 Fast (табл. 2).

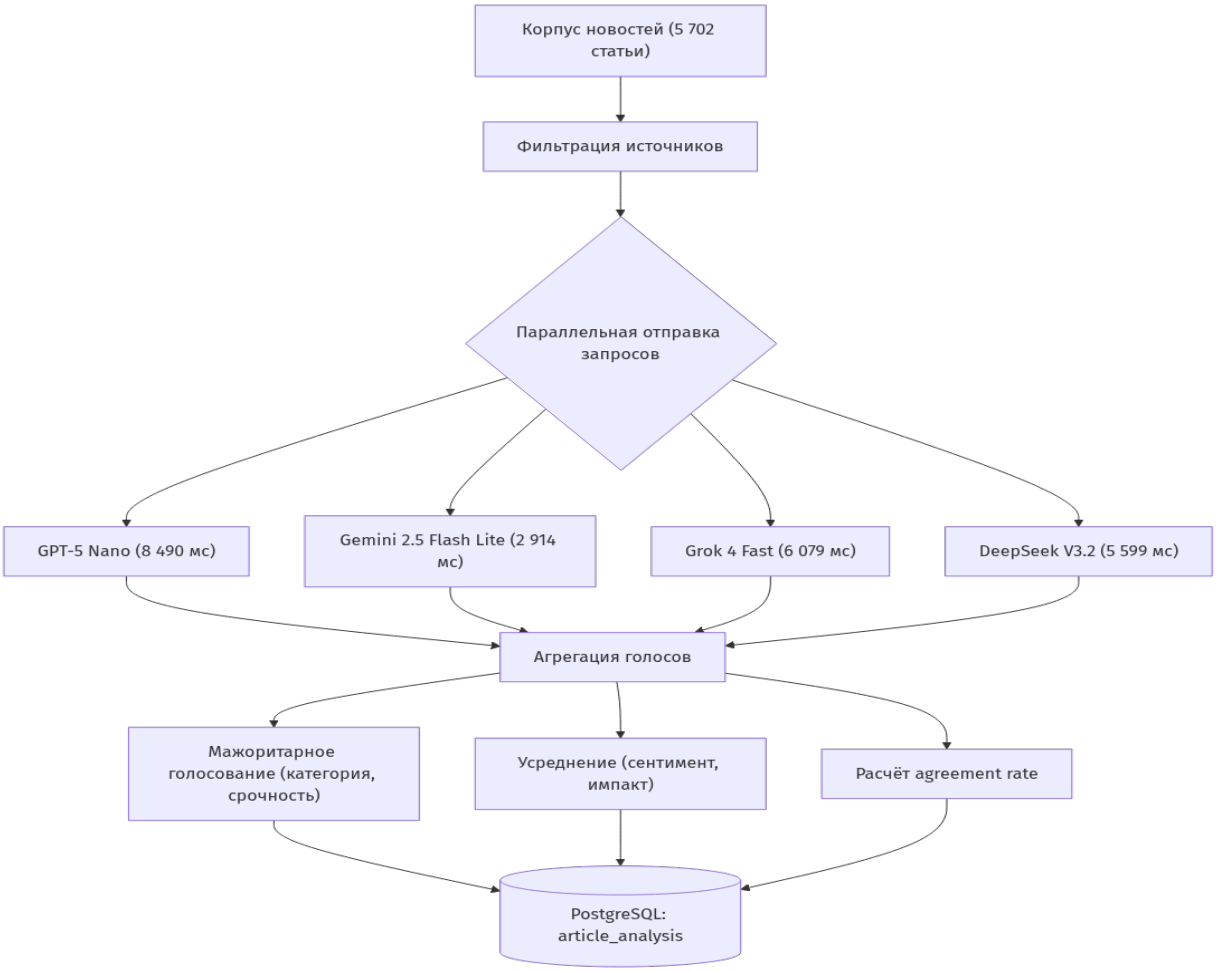

Рис. 1. Архитектура конвейера системы ELNA

III. Структура разметки

Каждый агент классифицирует документ по шести полям. Для тематической категории введена система кодификации K1K6, описанная в табл. 1. Тип срочности принимает четыре значения: срочно (Breaking), Запланировано (Scheduled), Прогноз (Prediction), Обзор (Recap). Тональность $s_i \in [-1{,}0; 1{,}0]$, значимость $p_i \in \{0{,}1, \dots, 10\}$.

ТАБЛИЦА I.

Код	Тематическая категория
K1	Регулирование и государственная политика
K2	Рыночная динамика и волатильность
K3	Технологии и кибербезопасность
K4	Внедрение и партнёрства
K5	Макроэкономика и геополитика
K6	Экосистема проектов

IV. Формализация метрики agreement rate

Пусть $S = \{s_1, s_2, \dots, s_k\}$ множество сентиментальных оценок k успешно ответивших агентов. Среднее значение:

$$\bar{s} = \frac{1}{k}\sum_{i=1}^{k} s_i \quad (1)$$

Среднее абсолютное отклонение оценок от среднего:

$$\sigma = \frac{1}{k}\sum_{i=1}^{k} |s_i - \bar{s}| \quad (2)$$

Согласно фундаментальным положениям статистического анализа [1], интерпретация дисперсии как меры отклонения значений от их математического ожидания позволяет рассматривать σ как индикатор консолидированности ответов ансамбля. Поскольку $s_i \in [-1; 1]$, теоретически максимальное значение σ равно 2. Метрика agreement rate определяется нормированным отклонением:

$$AR = 1 - \frac{\sigma}{2}, AR \in [0; 1] \quad (3)$$

При $AR = 1{,}0$ все агенты вернули идентичный сентимент; при $AR = 0{,}0$ оценки диаметрально противоположны. Интерпретация в рамках теории нечётких измерений [9]: величина $1 - AR$ является мерой неопределённости при отображении документа в пространство меток. Для категориальных полей итоговая метка определяется мажоритарным голосованием, выбирается значение, встречающееся у наибольшего числа агентов [5].

Документы с $AR < 0{,}85$ идентифицируются как нечёткие и направляются на ручную верификацию.

V. Описание корпуса и условий эксперимента

Корпус сформирован из криптовалютных и финансовых новостей источников, прошедших ручную верификацию на релевантность. Использовались четыре LLM-агента, запускаемые параллельно через REST API

при фиксированном системном промпте. В табл. 2 приведены технические характеристики агентов.

ТАБЛИЦА II.

Модель	Ошибок, %	Ср. задержка, мс	Ср. токенов ответа
Grok 4 Fast	7,84	6 079	55
DeepSeek V3.2	1,68	5 599	61
Gemini 2.5 Flash Lite	1,65	2 914	71
GPT-5 Nano	0,39	8 490	1 117

Из табл. 2 следует, что Gemini 2.5 Flash Lite обеспечивает наименьшую задержку (2 914 мс) при приемлемой частоте ошибок (1,65), тогда как GPT-5 Nano генерирует наибольший объём текста ответа (1 117 токенов) вследствие развёрнутого рассуждения. Наибольший процент ошибок зафиксирован у Grok 4 Fast (7,84) результат превышения лимита запросов.

Основные статистики AR по всему датасету приведены в табл. 3.

ТАБЛИЦА III.

Показатель	Значение
Среднее AR	0,9418
Стандартное отклонение	0,0350
Минимальное значение	0,6600
Максимальное значение	1,0000
Доля статей с AR 0,90	91,32
Доля статей с AR 0,85	98,54
Доля статей с AR 0,85	1,46
Доля статей с AR 0,80	0,35

Распределение AR имеет выраженный правосторонний пик (табл. 3): более 87,9 наблюдений сосредоточено в диапазоне [0,90; 1,00].

VI. Анализ agreement rate по тематическим категориям

В табл. 4 представлено сравнение метрики AR в разрезе тематических категорий (коды К1К6 расшифрованы в табл. 4). Категории отсортированы по возрастанию среднего AR.

ТАБЛИЦА IV.

Категория	Кол-во статей	Ср. AR	Ст. откл.
К1	580	0,9354	0,0371
К4	763	0,9357	0,0318
К6	1 200	0,9388	0,0336
К5	344	0,9391	0,0434
К3	183	0,9416	0,0315
К2	2 632	0,9468	0,0342

Наименьшее согласие зафиксировано для категорий К1 (Регулирование) и К4 (Внедрение и партнёрства) текстов, допускающих широкую интерпретационную вариативность. Наибольшее согласие достигается для К2 (Рыночная динамика), что объясняется структурированностью рыночных данных и однозначностью их тематического контекста.

VII. Корреляционный анализ

Для выявления факторов, влияющих на уровень межмодельного согласия, вычислены коэффициенты корреляции Пирсона. Результаты представлены в табл. 5.

ТАБЛИЦА V.

Пара переменных	Коэффициент корреляций	Характер связи
AR Impact Score	0,0664	Слабая положительная
AR Абс. значение сентимента	0,2852	Умеренная положительная
AR Кол-во токенов	-0,0915	Слабая отрицательная

Наиболее значимой оказалась корреляция метрики согласия AR с абсолютным значением сентимента (r 0,29). Полученные нами данные показывают, что тексты с выраженной (полярной) тональностью порождают более согласованные оценки ансамбля: эмоциональный сигнал однозначен для всех агентов, и разброс s_i минимален. Это не удивляет: модели сходятся на полярных текстах — сигнал ясен, и смысловой разброс снижается; нейтральные или смешанные тональности, напротив, путают классификаторы и порождают больше неопределенности (см. общие подходы в [3; 8]).

Отрицательная, но очень слабая корреляция с числом токенов (r 0,09) показывает, что объем документа почти не меняет уровень согласия. Длинные тексты дают побольше контекста, и да — это слегка увеличивает вариативность трактовок.

VIII. Верификация категориальных меток

Параллельно с мягкой метрикой AR проверили точность каждого агента в категориальной классификации — кто точнее, кто слабее; сравнивали с консенсусом ансамбля (табл. 6).

ТАБЛИЦА VI.

Пара переменных	Коэффициент корреляций	Характер связи
DeepSeek V3.2	95,47	93,74
Grok 4 Fast	88,92	87,78
GPT-5 Nano	77,78	81,29
Gemini 2.5 Flash Lite	70,24	88,80

Из табл. 6 следует, что DeepSeek V3.2 наиболее стабильно воспроизводит консенсусные метки (95,47 по категории). GPT-5 Nano, несмотря на максимальный объём ответа, демонстрирует наибольшее число категориальных расхождений (77,78), что указывает на высокую вариативность его рассуждений. Gemini 2.5 Flash Lite заметно расходится по категории (70,24). Зато по срочности согласие высокое (88,80). Похоже, одна задача требует более абстрактных суждений, другая — прямых меток.

IX. Анализ результатов и обсуждение

Высокий средний AR — 0,942 при стандартном отклонении 0,035 — говорит о воспроизводимости ансамблевой разметки: большинство документов получают согласованные оценки от четырех независимых агентов. Пожалуй, это дает основание применять мажоритарное голосование для автоматической проверки меток без ручной правки [6].

Из 5 702 статей лишь 83 (1,46) получили AR 0,85. Разбор этой горстки выявил три частых проблемы. Во-первых, минималистичные заголовки без смыслового контекста — например, avax или BTC Short Update — не дают агентам достаточно материала для согласованной классификации. Во-вторых, мультимедийный контент, особенно видео с коротким описанием, дает разную реакцию оценщиков. И в-третьих, прогнозы с

неоднозначной тональностью — скажем, Report: Tokenized Gold Will Trigger the Collapse of Global Paper Gold Market — легко открываются для разных толкований, что приводит к разбросу s_i.

Тип срочности *Запланировано* демонстрирует наименьшее согласие (AR 0,9327) предположительно из-за размытой временной отнесённости анонсируемых событий, которая по-разному оценивается агентами. Тип *Прогноз* показывает наибольший AR (0,9450) вследствие стереотипичности формулировок прогностических текстов.

X. Заключение

В работе формализована и экспериментально верифицирована метрика agreement rate как мягкая мера согласия ансамбля больших языковых моделей в задаче автоматической разметки новостных текстов. Метрика основана на нормированном среднем абсолютном отклонении сентиментальных оценок ансамбля и интерпретируется в рамках теории нечётких измерений как мера неопределённости результата разметки.

Основные результаты работы:

- Среднее межмодельное согласие по корпусу из 5 702 статей составило 0,942 при стандартном отклонении 0,035.
- 91,32 статей получили AR 0,90, что подтверждает высокую воспроизводимость ансамблевой разметки.
- Выявлена умеренная положительная корреляция AR с полярностью тональности (r 0,29): явно эмоционально окрашенные тексты анализируются ансамблем более согласованно.
- Предложен практический критерий качества: документы с AR 0,85 рекомендуется направлять на ручную верификацию; таких документов в корпусе 1,46.

Перспективными направлениями развития работы являются: расширение метрики AR на категориальные поля через нечёткую меру рассеяния голосов, апробация на корпусах смежных предметных областей, а также разработка адаптивного порогового критерия качества, зависящего от тематической категории документа.

Список литературы

[1] Машинное обучение: учебник / Е.Ю. Бутырский, В.В. Цехановский, Н.А. Жукова [и др.]; под ред. В.В. Цехановского. Москва: Директ-Медиа, 2023. 368 с.

[2] Araci D. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models // arXiv preprint arXiv:1908.10063. 2019.

[3] Bollen J., Mao H., Zeng X. Twitter Mood Predicts the Stock Market // Journal of Computational Science. 2011. Vol. 2, No 1. P. 1–8.

[4] Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D. M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 1877–1901.

[5] Cohen J.A Coefficient of Agreement for Nominal Scales // Educational and Psychological Measurement. 1960. Vol. 20, No 1. P. 37–46.

[6] Dietterich T.G. Ensemble Methods in Machine Learning // Multiple Classifier Systems. First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000. Proceedings / J. Kittler, F. Roli (eds.). Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. P. 1–15.

[7] Liang P., Bommasani R., Lee T., Tsipras D., Soylu D., Yasunaga M., Zhang Y., Narayanan D., Wu Y., Kumar A., Chen B., Koh P. W., Kapoor S., Narayanan A., Agrawala S., Ribeiro M. T., Saharia C., Zhou Q., Wang H., Zhou Q. et al. Holistic Evaluation of Language Models // arXiv preprint arXiv:2211.09110. 2022.

[8] Malo P., Sinha A., Korhonen P., Wallenius J., Takala P. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts // Journal of the Association for Information Science and Technology. 2014. Vol. 65, No 4. P. 782–796.

[9] Zadeh L.A. Fuzzy sets // Information and Control. 1965. Vol. 8, No 3. P. 338–353.